Magnar Brekke*

Popular vs. Professional Aspects of Economics Texts in English

Abstract

This article takes up the important issue of distinguishing popular from professional texts using economic texts as the object of analysis. It reports on a limited corpus study of authentic texts produced by native English professionals and published for native English readers, but differing in the nature of intended audience (lay/expert). Economics/electronic money transactions constituted the textual domain/subdomain, and the dominant text type was expository. Following a basic quantitative check and a tentative readability ranking the text samples were scrutinized with respect to lexical profile, frequency band penetration, terminological density and uniqueness. A detailed collocation study was then made of focal multi-word terms as well as personal pronouns. Standard corpus techniques were employed in exposing regularities and producing supporting documentation. At the same time possible reflections of communicative situation, pragmatic purpose and semiotic significance were noted. An attempt was then made to integrate such internal ("linguistic") and external ("situational") variables as revealed by a comprehensive textual analysis in a synopsis intended to bring out a possible clustering of features. The general question to which an answer was sought was to what extent is each perspective necessary and/or sufficient in defining the textual genre?

1. Introduction

In terms of the conventional (Bühler-type) "sender-content-recipient" triangle languages for special purposes (LSP) gravitate towards the content corner in being chiefly concerned with "Darstellung" of factual information. The resulting texts tend to fall into a definable set of textual genres appropriate to the domain and the ultimate textual function. However, the nature of the intended recipient will obviously influence the style of representing that factual information; thus in the analysis both text-linguistic, pragmatic and semiotic features can be

---

* *Magnar Brekke*
  Norges Handelshøyskole

seen to reflect whether the text is pitched at a popular or professional audience.

LSP comprehension has been shown to benefit considerably from readers' genre knowledge, i.e. from awareness of the repertory of conventional text forms and their functions in the domain community. Without opening up the Pandora's box of genre definitions one might ask whether such independent genre knowledge is necessary or at least strongly conducive to LSP comprehension, or whether the observable text-linguistic, pragmatic and semiotic features of a given text provide sufficient basis for readers to work out the message and indeed the genre for themselves.

The present paper reports an attempt to put this question to an empirical test. The preliminary conclusions are that while superficial statistical measures constitute helpful clues neither internal nor external criteria are in themselves sufficient. Nevertheless their clustering contributes substantially to an empirical basis for useful classification of professional genres. It is argued that such preliminary results have implications for corpus building, corpus studies, translator training and LSP teaching programs in general.

2.   Presentation of project

2.1. Material

A pilot study was undertaken of five authentic texts produced by native English writers and published for native English readers, but differing in the nature of intended audience (lay or expert). All five texts fall within the domain of Economics and more specifically under the subdomain of "electronic money transactions", as determined through a cursory reading. The dominant text type is expository, and the length of each text varies from 874 to 3776 words, a total of 8.834 words, the average text being approx. 1600 words long. The texts are complete texts rather than fragments, but for reasons not relevant to this paper one text has been shortened (by deleting whole subsections) to about 1/3 of the original. Four of the texts were available on paper in their published form, one was available in electronic form only. Two of the texts have been taken from professional journals, one from a newspaper, one from a popular science magazine, and the electronic one has been downloaded over the WWW.

2.2. Method

The text samples were scrutinized by the use of a standard concordance package (System Quirk). The general statistical picture emerging is summarized in table 1 below and then further elaborated. At the same time possible reflections of communicative situation, pragmatic purpose and semiotic significance were noted. Standard corpus techniques were employed in exposing regularities and producing supporting documentation.

2.3. Goal

The general question to which an answer was sought was: In view of the range of internal (primarily statistical and text-linguistic) and external (pragmatic and semiotic) variables revealed by a comprehensive textual analysis, to what extent is the respective type of criteria necessary and/or sufficient in defining the textual genre? More specifically: Given five short texts from the same subdomain representing the same general text type, is it possible to work "bottom-up" through textual/linguistic features to arrive at a clear generic typology as regards the type of audience which the author(s) may have had in mind, or are we dependent on "top-down" processing taking published form into account? The crucial test will be the electronic text, since it lacks the institutionalized publication form which normally answers such a question. Can we identify a text's genre "in the nude", so to speak, unassisted by clues from the dress it is wearing?

2.4. Overview of paper

After a brief survey of basic text statistical measures and an ad hoc readability ranking based on them the paper examines two lexical aspects of the text samples. One concerns the penetration of high frequency lexical words (appearing in all five texts) into the higher frequency bands normally reserved for function words, while the other picks up lexical words at the lower end of the frequency list which are either unique to one text or have a clear "tell-tale" function in (partially) revealing textual identities. It then focuses on important collocational patterns found with selected terms and with specific pronouns. Supplementary elements from the pragmatic and semiotic dimensions are pointed out, and a synoptic view of text profiles and feature clustering is provided. This leads directly on to a matching of each feature

cluster with its specific text, the identity of which has been kept from the reader up to this point. The relative success of this undertaking is claimed to have implications for corpus-related LSP research, teaching, and practice.

3. Findings: Internal variables

3.1. Texts statistics

Table 1 exhibits some important measures. Since the text samples are very small I make no claims on statistical significance or strict validity for these numbers, but they serve as a fairly concrete point of departure for uncovering aspects and dimensions which are worth pursuing.

Table 1: Basic measures of text samples

	1: WRD	2: CHAR	3: AWL	4: WPS	5: type/token
A	1625	8661	5.39	21.1	.45
B	1646	9146	5.60	25.9	.45
C	3776	19962	5.28	24.9	.45
D	913	4508	4.93	22.3	.46
E	874	4587	5.24	24.3	.48

The texts are randomly assigned letters A through E (left-hand margin of table 1). Columns 1 and 2 give word and character count, respectively, and columns 3 and 4 the average length of words and sentences. Text B has the longest words (5.60 characters), text D the shortest (4.93 characters). While differences in sentence length are relatively small, text B also has the longest sentences (25.9 words) while text A has the shortest (21.1 words). The type/token ratios in column 5, calculated for the first 850 words of each text (cf. Biber 1988: 239) reveal no significant differences among the text samples except for E, which has less lexical repetition than the others.

At least two of the three quick and superficial measures exhibited in table 1 point to relative differences holding among our five samples which can be brought together under a composite rank number: assuming that a text characterized by short words, short sentences and a high degree of lexical repetition (i.e. redundancy) is easier to read than one

with longer words and sentences with lower degree of repetition, I have allowed myself to assign "readability points" from 1 to 5 according to their rank in these respects (see table 2). Thus on a scale from 3 to 13 text D would appear easiest, B most difficult to read.

Table 2: Readability ranking

	1: WPS	2: AWL	3: type/token	4: Composite
A	1	4	3	8
B	5	5	3	13
C	4	3	3	10
D	2	1	2	5
E	3	2	1	6

3.2. Lexical aspects

Table 3 gives frequency ranks 1-10 (lefthand margin) of lexical words occurring most frequently in the five texts (A-E horizontally). The juxtaposition of these frequency lists creates a clear impression of sub-domain relatedness as far as shared lexical content (boldface items) is concerned, although the degree of relatedness varies.

Table 3: Top ten lexical word frequency tables

Rank	A	B	C	D	E
1	**bank**	financial	**electronic**	**card**	**cash**
2	**card**	service	**card**	Mr.	**electronic**
3	signature	software	issue-	**electronic**	**card**
4	key	customer	**money**	security	smart
5	**information**	**bank**	**cash**	**transaction**	system
6	digital	industry	**bank**	business	**transaction**
7	record	institution	value	product	**information**
8	organisation	on-line	**transaction**	world	**bank**
9	**electronic**	value	policy	years	marketing
10	privacy	player	scheme	currency	**money**

Taking the full spread of all words, lexical as well as function words, and then identifying the frequency bands where the eight most frequent lexical words turn up within the first percentile (1-100), gives a somewhat different perspective, see table 4.

Table 4: Penetration of top lexical items into general frequency chart

	card	elec-tronic	bank	cash	trans-action	money	secur-ity	infor-mation	TOTAL FRQ
A	13/20	8/40	15/20	6/50	3/100		6/50	10/20	61
B			13/20					7/50	20
C	62/10	66/10	27/30	28/20	16/40	60/20	11/60		270
D	20/10	6/30		4/50	6/30	2/80	6/30		44
E	12/10	22/10	5/30	22/10	8/20	4/40		5/30	78

Table 4 is to be understood as follows: the eight most frequent lexical words are listed horizontally across the table. For a given cell, the number to the left of the slash (/) gives the frequency of the relevant item in each text (lefthand column), while the number to the right of the slash gives the frequency band (1-10, 11-20, 21-30 and so on) in which the item appears. The lexical unit *card* thus occurs 13 times in text A and is found in the 11-20 frequency band. It will appear that the first five words (*card, electronic, bank, cash* and *transaction)* occur frequently in **four** of the texts; the next three words (*money, security,* and *information*) occur frequently in **three** of them. The lexical unit *transaction* is thus barely within the first percentile in text A but rises to the 31-40 frequency band in text C, the 21-30 band in D, and the 11-20 band in E.

This use of frequency bands constitutes a way of indicating how far a specific lexical word has "penetrated" into the higher frequency bands. Under normal distribution in a very large corpus the first percentile would be more or less "reserved" for function words. In the 28 million word Longman/Lancaster Corpus, for instance, only the words *man* and *little* reach the first percentile; i.e. of the first hundred most frequent words, 98 are function words (Brekke, Myking and Ahmad 1996). In an LSP corpus, on the other hand, words semantically related to the textual domain (i.e. terms) will be found much higher in the frequency charts.

What we are discussing here is a form of terminological density, and in order to arrive at an intuitive measure of this I have for each text calculated the percentage which these eight technical terms (which most texts share) constitute of the total word mass. Table 5 repeats (from table 1) the numbers for text lengths in column 1 and

(from table 4) the total frequencies of the eight most frequent lexical words in column 2. Column 3 expresses the percentages that the eight most frequent technical terms (given across top row of table 4) represent out of the total word mass for the respective texts.

Table 5: Terminological density

	1: WRD Text mass	2: TOT.FRQ 8 most frq.	3: % 2:1
A	1625	61	3.75
B	1646	20	1.21
C	3776	270	7.15
D	913	44	4.82
E	874	78	8.92

Not unexpectedly (given table 4) the total occurrence of the eight shared high frequency lexical words in text B makes up only 1.21% of the total word mass, counting all word forms, whereas for text E the corresponding percentage is close to 9. We interpret this to indicate that text B is not strongly focused on the shared domain topic while text E is. On this score text C is also strongly focused on common subject matter, while A and D are intermediate.

Equally interesting, of course, are the unique occurrences at the tail end of the frequency scale, lexical words occurring in only **one** of the documents and/ or with very low frequency; see table 6, which in addition displays "tell-tale" items giving strong clues as to textual genre.

Table 6: "Tell-tale", unique or low frequency items

	Lexical words/terms
A	digital/blind signature, untraceable, tamperproof, representatives, observers
B	on-line personal financial services, incumbent
C	issue, issuer, issued, issues, schemes, central bank
D	says, explains, believes, points out, denies, cryptography, algorithm
E	says, systems integrator, decipher, applets, prototyping, encrypted

The "tell-tale", unique or low frequency items given in table 6 appear to be just as revealing as to the genre identity of the respective text as the high frequency items in table 3. Text B in table 6 highlights the well-known limitation of simple frequency analysis, which counts only single words, i.e. character strings separated by spaces, and thus misses multi-word terms or expressions, by far the most significant lexical units in LSP texts. These can easily be brought out by producing a concordance for the relevant items. Thus the word *online* occurs in two of the texts under investigation, but only in text B does it form part of the expression *online personal financial services.* We will now examine some of the characteristic combinations which our selected terminological units enter into, and then look at pronoun occurrences.

3.3. Focal terminology

Figure 1 is a graphic integration of concordance information for each word in the phrase *online personal financial services* occurring uniquely i text B.

Boldface entries co-occur with any listed word bordering on their territories; other entries co-occur with listed words on their corresponding level. It will appear that the boldface expression is fairly rich in its collocational patterns. Some of these undoubtedly qualify for term status, but that is a different undertaking.

Figure 1: Key collocations, text B: online personal financial services

The expression displayed in figure 1 was unique to text B. At the opposite extreme we find the phrase *electronic bank notes/cash/money*, which occur in all texts except B. Figure 2 displays the concordance information for this phrase while figure 3 does the same for *card/cards*.

Figure 2: Key collocations, texts A & C-E: electronic bank notes

		commerce forms	consultancy
director of		**cash**	
Internet-based smartcard based denominal untraceable	e- encrypted	purse coins value	systems offers transactions
		money	products systems
			chip laundering schemes services
	national central		account
the	**electronic**	**bank**	**notes**
first real	digital		branch networks
	cross- lg.qt.ies of counterfeiting of		
		credentials documents privacy replacement signatures	
to authenticate achieving			
	blind forged		

Figure 3: Key collocations, texts A & C-E: card(s)

a single- limited multi- global universal	major purpose function	credit or debit chip smart telephone pre-paid plastic			number issuer limits reader
	currency				
	voter	registration	**card(s)**	-based	schemes microprocessors
	contactless				transactions
genuine other	Mondex AmXpress Visa 'micro'- magnetic	loyalty check guarantee payment			industry
a	massive	stripe			player

3.4. Pronouns

Table 7 displays the relative incidence of all pronouns (column 1) in the five texts and further specifies percentages of individual items; it also lists those that are absent or prominent (first and second column from the right, respectively) in a given text.

Table 7: Pronoun incidence

	% of all words	% of all pronouns						Prominent other than *it*	Absent
	pronouns	1st	2	3m	3f	3it	3pl		
A	3.80	3.2	13	1.6	11	32	39	you she	
B	2.24	0	0	0	0	46	54	they	I you he she
C	1.56	0	0	3	4	57	36	they	I you
D	4.81	44	7	12	0	30	7	I he	
E	2.28	25	5	5	0	60	5	I he	

It will appear that of the total word mass in each text, D has the highest proportion of pronouns (4.81% of text mass is pronouns, chiefly *I* and *he*), whereas in C only 1.56% of all words are pronouns, all third person). B appears to be an impersonal text with only *it* and *they*. We will now take a closer look at the characteristic verb phrases taking these pronouns as subject. Figures 4, 5

and 6 exhibit the collocations of first, second and third person personal/possessive pronouns or names, respectively.

Evidence from pronoun collocation (as displayed in figures 4-5-6) and similar patterns has immediate communicative implications, since first person pronouns reflect subject or agent orientation, while second person pronouns reflect object orientation. First person pronouns as well as third person pronouns and names are often associated with statements in a reporting context such as news, which is beginning to suggest itself as a likely genre for text D. However, we still need more clues before definite genre identifications can be made, for which we will now turn to external situational variables manifested in the five texts.

Figure 4: First person collocations

Text	Pronoun	Verb phrases
A		am associated call a computer (have) developed/demonstrated/proposed/shown
D	**First pers.**	am delighted/delirious decided expect said think started leaping about
E		can then go to a are doing as well/better are very serious/driving Mondex asked ourselves found no business have cracked the problem/a powerful start/spent a million/ to account for

Figure 5: Second person collocations

Text		Pronoun	Verb phrases
A	what whom every time where	**Second pers./** **(poss.)**	buy communicate with make a telephone call travel life/medical history/taxes
D			can't trust have to account for wouldn't worry shan't

Figure 6: Third person collocations

Text	Verb phrases	Pronoun name	Apposition	Verb phrases
D	explains says points out	**Mr. Jones**		had learned recognized believes denies
		he		had spent time has signed up says suggests
E	points out says comments	**Mr./David Birch** **R. Nuttall** **Harvey** **Mike Nash**	Director of Manager of Vice President of	

4.   Findings: External variables

The relative incidence of pronouns has already given some hints as to the pragmatic context of some of the texts and inferences one might draw from the occurrence of *I/you*. In this section we will briefly expand this perspective to include other aspects of the situational and semiotic elements manifested in the text samples; we will take the texts one by one.

4.1. Text A

This sample contains an invented story illustrating the operation of electronic signatures. The characters appearing in the (clearly didactic) story are Alice and Bob, prominence being given to *she/her*. Most of the *you*-forms appear in the initial "attention-grabber".

4.2. Text B

Interspersed in the running text are illustrations in the form of tables and diagrams as well as enhanced quotes repeating key statements. Personal pronouns are totally absent.

4.3. Text C

This text also has pie-charts and diagrams interrupting the flow of text.

4.4. Text D

As noted above, this sample has the highest pronoun incidence of all, and half of them are *I/you*. The item *Mr.* appears half-a-dozen times, direct quotes abound. Verbs with first person subject are mostly emotive or indicating mental activity, those with third person subject often cognitive (cf. table 6 and figures 4-5-6). A URL is given below the text.

4.5. Text E

Mr. appears once, direct quotes abound, text is sprinkled with self-evaluative verbs, company names, and personal names with titles in apposition (cf. table 6 and figures 4-5-6). E-mail address is given.

5. Text profiles and clustering

We have now reached the point where some of the varied measures and scattered observations made above may be brought together synoptically in an attempt to address the general dichotomy posed in the title: Are there aspects of the text samples which reflect popular or professional language use, which can be used in further identifying the genre of a given text? Table 8 integrates information given above in a synoptic view:

Table 8: Synopsis of features and observations

	Read-ability	Termin. density	Pronouns: incidence	Pronouns: prominent	Pronouns: absent	Pragmatic & semiotic
A	8	3.75	3.80	you she		Invented story
B	13	1.21	2.24	it they	I you he she	tables diagrams enh. quotes
C	10	7.15	1.56	it they	I you	piecharts diagrams
D	5	4.82	4.81	I he		direct quotes URL
E	6	8.92	2.28	I he		definitions direct quotes e-mail

The question facing us at this point is the following: Within the sub-domain of "electronic money transactions", are there consistent lin-

guistic criteria, supplemented by pragmatic and semiotic hints, which give us a clue to the textual genre being sampled?

Text A in this perspective has medium readability, medium terminological density, but high pronoun incidence. Notable among pronouns used are *you* (reflecting attempts to engage the reader) and *she* (arising from the embedded didactic story). This is clear evidence of a popular genre, and accords well with the overall impression created by articles published in *Scientific American*, which is the source of text A (Achieving Electronic Privacy [abbreviated], by David Chaum.

Text B marks something of a contrast to A in that it has very low readability (scoring all the 13 possible points, see table 2), very low terminological density, and medium pronoun incidence. Here *you* and *she* (and the remaining personal pronouns) are conspicuously absent, yielding the arena to the impersonal *it* and *they*. The running text is supplemented by tables and diagrams and punctuated by enhanced quotes. Readers familiar with *The McKinsey Quarterly* will perhaps recognize the journal's style of presentation, well illustrated and yet formal without inundating the reader in technical terminology. The genre is thus that of a professional article ("Who will capture Value in On-line Services?" by Tab Bowers and Marc Singer) without being heavily academic.

Text C is similarly illustrated with charts and diagrams, which in itself would immediately raise a suspicion that we are looking at another professional journal article. The impersonal pronouns *it* and *they* are absolutely dominant, *he* marginal and *she* only used to achieve political correctness in the phrases *he or she* and *his or her,* while *I* and *you* are totally absent. Pronoun incidence is in fact lower in C than in any other sample, while its terminological density is second highest and readability second lowest. If there is an academic professional article among the samples, this has got to be the one - which appeared in the *Bank of England Financial Stability Review* as "Electronic Money", written by Mark Robson.

Text D puts us squarely back in the popular genres again, with top score for readability, medium score for terminological density and top score again for pronoun incidence, all of them, in fact, except *she*. First person *I/we* dominate, *it* is a strong second. Text D is peppered with direct quotes in the running text, which accounts for the consis-

tent first person focus and strongly suggests a reporting of statements. Text D is typical of feature articles appearing in the IT/Management section of the *Financial Times,* this one entitled "Vision for a global multi-currency card" and contributed by Rod Newing.

That leaves us with Text E, which is really the crucial test of our approach, since it was downloaded from a WWW-site without the normal genre-characterizing publishing channel. Its readability is almost as high as D, while its terminological density is much higher, in fact the highest of all samples. Pronoun incidence is among the lowest, but here *it* accounts for 2/3 of all pronouns, *I* for 1/4. The definitions occurring throughout the text are clearly provided as a didactic softening of the impact of the terminological density already noted, while the direct quotes again indicate a reporting function. The text was found by searching for key words (cf. table 3) on AltaVista, and it was only at a later stage that the present writer discovered that E had in fact been written by the same author as text D and indeed had appeared in the same newspaper three months earlier: "Electronic cash", by Rod Newing, in the *Financial Times*.

6.   Main claims and a conclusion

• Statistical footwork (provided by standard word processors and concordancing packages) is useful in text-based genre studies.

• Frequency studies give useful hints as to density of domain-focal lexical words (i.e. terms).

• Concordance studies reveal collocational patterns pointing to potential multi-word terms for the domain.

• Text-linguistic, pragmatic and semiotic features provide essential clues to function and purpose of text in domain community.

Conclusion: The pilot study of five short LSP texts reported here indicates that the specific question posed above can be given a positive answer: Yes, it is possible to work "bottom-up" through textual-linguistic features (supplemented by pragmatic/semiotic observations) to arrive at a generic typology of text forms. No feature in isolation but rather the clustering of features from various levels of analysis indicates significant convergences and contributes to placing genre taxonomies on a stronger empirical foundation.

The implication is that we should build, extend, and maintain LSP corpora for more domains, develop tools and methods for studying them efficiently, and establish and cultivate professional networks linking the expertise of text linguists, technical communicators, translators, LSP teachers and domain specialists.

Bibliography

1. Text samples

Bowers, Tab and Marc Singer (1996). Who will capture value in On-line Services? *The McKinsey Quarterly*, n°. 2, 78-83.

Chaum, David (1992). Achieving Electronic Privacy [abbreviated]. In *Scientific American*, 96-101.

Newing, Rod (1996). Electronic cash, WWW.

Newing, Rod (1997). Vision for a global multi-currency card. In *Financial Times*, 5 february.

Robson, Mark (1996). Electronic Money. In *Bank of England Financial Stability Review*, 23-30.

2. References

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Brekke, M, Myking, J. and Ahmad, K. (1996). Terminology management and lesser-used living languages: A critique of the corpus-based approach. In *TKE'96: Terminology and knowledge engineering*. Frankfurt/M.: Indeks Verlag.